InFoMM
Industrially Focused
Mathematical Modelling

University of Oxford

Machine Learning Based Triage Models for Remote Detection and Care of Influenza

Anna Berryman

Vironix is a telehealth monitoring company interested in providing at-home diagnostic tools to help inform patient decisions. It has developed CovidX, an app that uses machine learning to help patients decide, while at home, whether they need hospital care or can stay at home to recover. CovidX can advise patients on whether they have a severe or non-severe presentation of Covid-19. The Covid-19 pandemic has shown that it is imperative for people to be able to quickly assess their condition to help prevent the spread of the virus, and avoid unnecessary hospital visits.

To train machine learning models, Vironix needs patient profiles. It generates patient profiles from data available from past clinical studies, where data is presented in summary tables such as Figure 1. It then trains various machine learning classifiers to classify the severity of a patients' condition. In this project, we further develop data generation techniques and consider them in a novel application to influenza. We extend the work carried out by Vironix by considering different generation methods for creating patient profiles and test the different methods by using a real data example. This allows us to give a more thorough analysis of the different generation methods.

The summary tables found in the literature contain information on patient demographics (age and gender), symptoms (e.g. cough, headache, and shortness of breath), and comorbidities (e.g. liver disease, diabetes, and morbid obesity). We have the influenza data from various clinical studies and combining three influenza studies results in 26 characteristics for each patient profile.

There are multiple modelling choices we can make when generating data from the statistical tables. The first method we develop is the most simple; we create patient profiles simply by treating all characteristics as independent (termed the independence method). However, this results in unrealistic profiles such as patients aged 0-14 who smoke. In its CovidX paper, Vironix aims to create only realistic patients by incorporating external data regarding relationships between characteristics. In the method Vironix uses, we treat all characteristics as independent but then roll a dice to accept or reject a patient profile based on the characteristic relationships (termed the rejection method). Finally, we also consider a third method that uses a mathematical object called a copula to try to retain the distribution of characteristics from the influenza data, while also incorporating the characteristic relationships (termed the copula method).

We could compare various machine learning classifiers trained on data generated using these methods however, because we do not have patient data, we cannot assess how good these models are at classifying real patients. Therefore, instead of testing the methods against influenza data generated from summary tables, we will use real data from a slightly different situation to test the methods.

We consider data collected from March 2014 to July 2017 from one academic and two community emergency rooms in the U.S. All adult emergency room visits that resulted in either formal admission or discharge were recorded. The data contains raw patient data with a corresponding triage label, admit or discharge. We can now produce a statistical summary table similar to the one in Table 1. We then generate three data sets using the three methods explained above, and train three models, each using a different data set. We can now test the three models against real patient data, giving us a more realistic picture of how the model would perform when deployed in a real world scenario.

UKRI
Engineering and
Physical Sciences
Research Council

Table 1 – Partial table with patient demographics and clinical characteristic of 511 influenza patients in India

Characteristics	Severe influenza A (H1N1) (N=140) N(%)	Non-severe influenza A (H1N1) (N=371) N(%)
Cough	137 (97.9)	364 (98.1)
Fever (≥ 37.5°C	132 (94.3)	353 (95.4)
Sore throat	74 (52.9)	232 (62.5)
Shortness/difficulty in breathing	99 (70.7)	263 (70.9)
Nasal catarrh	48 (34.3)	136 (36.7)

Table 2 – AUC scores for four models tested on real data. Three trained on the three generated data sets and one trained with real data.

Method on which the model was trained	AUC
Independence	0.7314
Rejection	0.7307
Copula	0.7306
Real data	0.7409

Interestingly, we find that the results for each data generation method are very similar. We see the AUC scores (area under receiver operating characteristic curve score, similar to the accuracy of a model but it accounts for the imbalance between the severe and non-severe classes; a random model would have an AUC score of 0.5 and a perfect model would have an AUC score of 1) for these models in Table 2. The inherently simple and sparse data, mainly yes/no characteristics, could be the reason for this. This implies that, for a data set with a relatively small number of features (28), models trained with the data from different generation methods perform similarly to one another on real data. We can also train a model on the real data, from real patients. The classifier trained using this data performs similarly to those trained on the three generated data sets.

Summary

We developed three methods to generate patient profiles from summary tables presented in influenza studies. The most simple of these treats all symptoms and comorbidities as independent. The second aims to remove potentially unrealistic patients by rejecting patients based on external data, thereby skewing the marginal distributions for those characteristics with external data. The third aims to retain the marginal distributions while also incorporating the external data.

Each method has advantages and disadvantages, which we were able to test using real data. This test is important for Vironix to asses its data generation methods. Interestingly, the three methods for data generation performed similarly when tested on real data. This suggests that the data generated can be used to train models for classifying real data, even if not all patient profiles generated are perceived to be realistic. Sumanth Swaminathan, Co-founder and Director of Vironix Health, said:

> *"Anna's work began as a project to apply the quantitative methodologies that Vironix developed for Covid-19 to a class of flu-like illnesses that have prevailed and mutated around the globe for hundreds of years. Specifically, her task was to create a prediction model that could take human health data generally available at home to consumers (symptoms, simple physiological readings, demographic and profile info, etc) and classify scenarios that represent severe and/or non-severe presentations of influenza. Anna has done an excellent job in reading and catching up on the clinical literature to learn more about the underlying disease under investigation. She provided significant peer review to the portion of our methods that involves generation of patient scenario data for training prediction models. She further developed a new approach generating patient data, and she has compared our approaches in a thoughtful and methodical way. Anna has also built early prediction models for classifying influenza severity from consumer data. In general, I think that Anna has demonstrated strong acumen for modelling and research. She's learned new programming and data science hard skills in a short period of time. She has provided both critical review and new material for Vironix to leverage for future modelling efforts. Finally, she's given us a promising start to our flu prediction products. Anna is a patient communicator. She is talented, self-sufficient, hard-working, and curious... a great blend of skills to fuel a promising future."*